Why Professional Judgment Is Better Than Objective Description in Dental Faculty Evaluations of Student Performance


Abstract: Practices intended to increase the appearance of objectivity in grading may work at cross purposes with professional judgment. In this study, an analysis of two removable prosthodontics technique projects in one U.S. dental school found that the use of component criteria (checklist) grading was less consistent than overall judgments of the same work and less predictive of dental students' future learning. A factor analysis revealed latent structures in both projects that would make it inappropriate to use a component criteria approach for grading. Common defenses of objectivity—such as scientific foundation, the relationship between reliability and validity, and legal requirements—are questioned in this article, and it is shown how simple adjustments to judgment scores can be made more effective than checklists, faculty calibration, or deselecting faculty members and with better measurement and teaching features.

Dr. Chambers is Professor of Dental Education, University of the Pacific Arthur A. Dugoni School of Dentistry; and Dr. LaBarre is Associate Professor of Integrated Reconstructive Dental Sciences, University of the Pacific Arthur A. Dugoni School of Dentistry. Direct correspondence to Dr. David W. Chambers, Arthur A. Dugoni School of Dentistry, University of the Pacific, 2155 Webster Street, San Francisco, CA 94115; 415-929-6438; dchambers@pacific.edu.

Keywords: dental education, assessment, grading, removable prosthodontics, technique learning, faculty consistency, objectivity, professional judgment, calibration

Submitted for publication 10/18/13; accepted 12/24/13

The concept of “objectivity,” which we define as indisputable description, has enjoyed a seldom-questioned place of honor in dental education and practice that it may not deserve. By this definition, objectivity is more of an inquiry stopper than a place to start learning. This article will examine just what objectivity can and cannot do compared to the alternative of professional judgment. We will use the phrase “professional judgment” to refer to those qualified by training and experience using a comprehensive perspective to choose or endorse an action upon which they stake their reputation.

“Competence” is the capacity to perform appropriately in context, supported by relevant understanding and values.1-3 “Learning” refers to all the relatively permanent changes in that capacity that result from experience. Both competence and learning are inferred from observing performance. There is a danger in equating observed performance (which can be described) with capacity to perform (which must be inferred). Further confusion is introduced when objectively describing the product or process of performance without considering the context in which or the means by which it was produced. The danger is commonly known as “measurement error.”4

In the context of dental education, we are primarily interested in capacity.5 We want to know what students are likely to do the next time. We are especially keen to be accurate in predicting how they will perform in clinic based on what they do in lab and how they will perform in practice based on their performance in school. Certification for graduation is a promise to the public about capability. It is based on professional judgment. A score in a preclinical technique course, even if it perfectly describes the average of past performance, is only useful if it helps students become better dentists and assures other clinical faculty members that students will not require excessive instructional time or endanger patients.

Dental education often supports measurement rituals regarding the way scores have been determined and then accepts the protocol as evidence that the marks are reliable and valid. In this article, we will argue that these rituals have largely symbolic value and that what really matters are the decisions that flow from the evaluation. In particular, we will advance a view that all differences in evaluation score that are not used for decisions or that fail to improve future behavior are false precision and wasted effort at best. A score on the National Board Dental Examination (NBDE) sufficient to qualify for licensure or to get into graduate school is mostly what students care about. Few practitioners actually remember their overall score, let alone their scores on sections they
passed years ago. The difference between 69 and 72 on a classroom test does not matter unless 70 is the cut-off for something important such as a grade. If the cut-off score is changed to 75, the difference no longer matters. Fierce battles arise whenever the cut-off score is changed. The battles are about finding enough points to change a decision, and an objective review of the entire performance would be out of the question. Protests in the other direction over receiving a 97 when the student believed 95 was deserved (being perfectly objective) are unheard of. Measurement issues are not about accuracy in describing the real and objective nature of the performance. They are about getting the decisions right about capacity. They are about putting students in the right categories when different categories of students will be treated differently in the future.

The problem of diminishing potential disputes over objective descriptions is a real concern, but it is a secondary one. There is a different set of rituals to protect faculty members from having to defend their evaluations. Research on measurement error where faculty members evaluate work products such as preclinical or clinical test cases saw its peak in the 1970s and has since almost disappeared. This research showed that faculty consistency, both intrarater and interrater, is in the $0.200 < r < 0.700$ range.\(^{6-15}\) Calibration exercises, training on criteria, and fiddling with the scale or adjusting the weights could move the $r$-values up by a small, and occasionally significant, amount within this range. It is doubtful that a manuscript repeating this many-times demonstrated picture would be accepted for publication today. Our PubMed search of the literature failed to find any such studies in dentistry in the past twenty years and none at all demonstrating a lasting effect of modifying faculty ratings systems for projects. There are also no reports of differences between measurement methods and impact on learning or decisions elsewhere in the educational program and no studies that found the calibration consistency “too low.”

It may be fair to say that dental schools have accepted the middling range of consistency for faculty evaluations of student performance on projects and have turned their attention to matters of efficiency and defensibility of these practices. Innovations seem to be prompted by changes in course director, and reports now place consistency of faculty evaluation to be prompted by changes in course director, and defensibility of these practices. Innovations seem to have turned their attention to matters of efficiency and impact on learning or decisions elsewhere in the educational program and no studies that found the calibration consistency “too low.”

Despite lack of evidence for its effectiveness, the practice of faculty members evaluating projects by judging component criteria and then adding up the part scores has become a fixture in preclinical technique laboratory courses in dental education.

The evidence is inconclusive regarding the claim that faculty members are more consistent when basing their evaluations on summed parts of a project than when evaluating the project holistically.\(^{36-40}\) Judging the entire performance or product without specifically scoring component parts is sometimes referred to as the “glance-and-grade” procedure in the literature. There is only one study, a 1980 article by Goepferd and Kerber,\(^{41}\) that reports an advantage for summing of component criteria scores compared to overall professional judgment scoring.
It is frequently defended as being “more objective” and more “fair” than the alternative of asking faculty members to give their professional judgment regarding the overall quality of the project. The issue is what the American philosopher John Dewey labeled “warranted assertability”: what is true is what we can say without fear of being contradicted.55

The purpose of our research reported here is to explore the part-whole relationship that criteria bear to the overall quality of a project. Specifically, two questions will be probed. First, are features of a project independent of each other? It is an assumption of a component criteria grading system that the elements are independent and can therefore be added without danger of double counting. If there is an internal structure in the project, such an assumption will not hold, and a criterion evaluation system will be trading higher “objectivity” for lower validity. Second, are there characteristics of a project when viewed as a whole that cannot be reflected appropriately in a component criteria system? If so, there is reason to believe that comprehensive professional judgment is a more valid method of evaluation, even if not a more consistent one. In the discussion section, comments will be offered on evaluation for the purpose of description compared with evaluation for the purposes of diagnosis and prognosis. The prospect remains that the objectivity of evaluation can be raised to a point where it is defensible against criticism, but its value in the education program will have been diminished in the process.

Materials and Methods

This project was approved by the Institutional Review Board at the University of the Pacific in the expedited category, Protocol #10-33.4. Data were gathered from faculty ratings, student self-assessments, and student peer ratings involving two laboratory projects in a preclinical laboratory course in removable prosthodontics technique at the University of the Pacific Arthur A. Dugoni School of Dentistry. One hundred fifty-seven students (including twenty-seven international students taking the course as an integral part of the second-year dental student program) participated. At the end of the first quarter, students completed a denture set-up project in a test-case format. Six months later they completed an immediate denture project under the same conditions. When they had finished each project and before submitting it for faculty evaluation, students marked their own projects according to the designated criteria. After evaluating their own immediate denture projects, the course director assigned students to perform an anonymous evaluation on a project completed by a randomly chosen classmate using the same evaluation form.

For grading purposes, Projects were scored by a team of nine faculty members who also served as instructors in the lab. The typodonts were scored without knowledge of students’ identities. Each instructor scored a convenience sample of projects. Subsequently, the course director scored all projects before returning the typodonts and evaluation sheets to the students the following week.

Additional data were gathered to assess consistency of faculty evaluations. On each graded project, nine to thirteen typodonts were selected for immediate reevaluation by one randomly selected additional faculty member. This was called the “Field Consistency” exercise. For the immediate denture project, in addition to the Field Consistency check, the course director chose nine typodonts to represent a typical range of student performance, and each was evaluated by all faculty members. This was called the “Test Case Consistency” exercise. Thus, there were three measures of overall faculty consistency in the ratings: instructors compared with each other on randomly selected student cases as part of the grading exercise (Field Consistency), instructors compared with each other on a set of nine cases used for a calibration test (Test Case Consistency), and instructors compared with the course director.

Once each quarter, faculty members were asked to complete a two-item survey measuring where they felt they would stand compared to fellow faculty members as being a “hard” or “easy” marker and whether their marks would be more or less consistent than those of their colleagues. A seven-point Likert scale was used, and the average was taken over the three (sometimes two because of absences) self-reports for each faculty member. A score of 4 was typical. Higher scores indicated a self-perception of being an “easy” grader and a consistent or well-calibrated grader.

The forms used for evaluations were the same for students and faculty members, but faculty members did not have access to the students’ assessments. The denture set-up project was evaluated on seven criteria using a 0 to 3 scale (sometimes 0 to 2), with a maximum possible summary score of 20 points. The maximum point value available for the immediate denture project was 25, and eighteen dimensions...
were appraised on the evaluation sheet, with 0 as the lowest possible score and 1 or 2 the highest for each dimension. In most cases, the total score was the sum of the scores for individual criteria—but not always.

Both students and faculty members liberally used “half points,” marking between scale values, and gave a total score that sometimes differed from the numerical total of criterion points. In order to make meaningful comparisons between the two projects with differing total point values, the scores were converted to percentages and expressed at numbers between 0 and 100.

Overall project scores were compared using correlation coefficients in the usual manner. For component criterion scoring, where effective use of subscors was investigated, a different metric was required because evaluators used multiple characterizations of each project. For example, two faculty members could have agreed on an overall score for a project, and each identified exactly three criterion flaws. However, each may have noted different errors. Alternatively, a student who marked every questionable criterion as an error would have accurately found all flaws identified by faculty members but would not be considered an “accurate” self-assessor because he or she had failed to differentiate features that met criteria.

Average false positive and false negative scores were used to describe the accuracy of the component criterion scoring. First, cases in which the faculty member identified failure to meet criterion were identified. A flaw was defined as any mark other than the top possible score, with half scores also counting as flaws. The proportion of flaws noted by faculty members that were also noted by students was called the “Confirmed Flawed” percentage. The proportion of criteria marked by faculty members as acceptable where the student also marked the same criteria as acceptable was called the “Confirmed Acceptable” percentage. A low score on this measure reflected situations in which students thought they had made a mistake but faculty members felt the work fully met criteria. These two measures are somewhat independent of each other. A student could guarantee a perfect Confirmed Acceptable score by marking every criterion as having been met, but he or she would then have a low Confirmed Flawed score.

There were Confirmed Acceptable and Confirmed Flawed scores for each evaluation: one each for the denture set-up, the immediate denture, and the peer assessment of the immediate denture project. In each case involving students, the faculty score was considered the standard. Calculating a component criterion score for faculty members was more difficult because there was no “standard.” A boot-strapping technique was used for the cases involving faculty members. In this technique, one faculty member was arbitrarily selected as the standard, and false positive and false negative rates were determined for each of the other faculty members. Then the next faculty member was used as the standard and the procedure repeated until each faculty member had served as the standard. The scores for faculty members were taken as the average of their consistency with every other faculty member serving as the standard.

The procedure used to determine component criteria scores produced a two-by-two matrix that permits calculation of the chi-square statistic and the phi-coefficient. Phi is a measure of association ranging from 0.00 to 1.00 and is directly comparable to the r-value of the standard correlation coefficient. The latent structure of the dimensions (criteria) of the two projects was investigated using factor analysis. Faculty marks for each project were factor analyzed using a principal components technique and varimax rotation, with factors selected for interpretation determined by eigenvalues above 1.0 and by examination of the scree plots.

### Results

**Faculty Member Consistency**

Table 1 summarizes the characteristics of faculty members’ evaluations. The interrater consistency for the three measures of faculty calibration was remarkably similar, and the r-values are all in the modest range of the mid-0.50s. The results of the Field Calibration (pulling random cases from the grading session for double marking) agreed with the Test Case Calibration exercise for the immediate denture project, r=0.760, p<0.01. The course director was modestly calibrated with the rest of the faculty considering all projects, but not for the Field Calibration grading of projects r=0.139, p=ns. Faculty members who self-reported that they were “easy” graders tended to give higher scores on average (r=0.334, p=ns because of the small sample size) and to have fewer of the errors they found in student work confirmed by their colleagues (r=-0.338, p=ns). Those who self-reported that they would be consistent with colleagues tended to be more consistent for Field Calibration consistency (r=0.330), for Test
project, faculty members found flaws in an average of 3.7/7.0 criteria (52 percent) per project, while students faulted themselves on 3.3/7.0 (47 percent) of criteria. Although there was some degree of overall agreement, consistency in the details was much diminished. The phi-coefficient expressing consistency across criteria was only 0.194. This is significantly less than the $r=0.262$ for correlation using overall scores ($p=0.05$ by test for differences in strength of association). On the second project, faculty members found errors on an average of 17 percent of the criteria and students found errors on 15 percent of the criteria. But there was almost no agreement about where the offending characteristics were ($\phi=0.089$). This is substantially less than the $r=0.407$ for agreement on overall score ($p<0.001$ for difference in associations). The detailed analysis by criterion for peer assessment showed virtually identical error rates for students and faculty members and very weak agreement.

The results of the bootstrapping exercise with faculty members for identification of specific criteria rather than overall scores are also shown in Table 2. The $\phi$-statistic is 0.375. This means that faculty members agreed among themselves on which specific criteria were troublesome better than they agreed with students ($\phi=0.375 > \phi=0.089$, $p<0.01$). However, they were less consistent when using specific criteria than they were when they made an assessment using an overall approach based on professional judgment ($\phi=0.375 < r=0.575$, $p<0.01$).

### Table 1. Consistency of faculty ratings of two projects in preclinical removable prosthodontics technique laboratory (N=10 faculty members [9 row instructors and the course director] and 154 students)

<table>
<thead>
<tr>
<th>Consistency</th>
<th>Mean</th>
<th>SD</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field consistency</td>
<td>0.575</td>
<td>0.132</td>
<td>0.760</td>
<td>0.139</td>
<td>-0.391</td>
<td>-0.068</td>
<td>0.646</td>
<td>0.298</td>
<td>0.330</td>
</tr>
<tr>
<td>Test case consistency</td>
<td>0.565</td>
<td>0.185</td>
<td>0.461</td>
<td>-0.020</td>
<td>-0.235</td>
<td>0.001</td>
<td>0.171</td>
<td>0.549</td>
<td></td>
</tr>
<tr>
<td>Course director consistency</td>
<td>0.577</td>
<td>0.153</td>
<td>0.457</td>
<td>-0.362</td>
<td>0.139</td>
<td>0.090</td>
<td>0.501</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confirmed Acceptable %</td>
<td>85.8%</td>
<td>4.9</td>
<td>-0.849</td>
<td>-0.158</td>
<td>0.137</td>
<td>0.232</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confirmed Flawed %</td>
<td>38.9%</td>
<td>11.6</td>
<td>-0.025</td>
<td>-0.338</td>
<td>-0.326</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Faculty views of own performance |

| 6. Average score given           | 82.4   | 3.53  |       |       |
| 7. Self-reported easy grader     | 4.1    | 0.702 |       |
| 8. Self-reported consistent grader| 4.2   | 0.531 |       |

Case Calibration consistency ($r=0.549, p<0.05$), and with the course director ($r=0.501, p=0.06$). They also had fewer of their flaws confirmed by colleagues ($r=-0.326, p=ns$), and they self-reported that they were “easy” graders ($r=0.584, p<0.05$).

A large part of the difference among student scores is a function of which faculty member performed the evaluation. The standard deviation of scores across faculty members was 3.5 on a scale that ran in practice between 50 and 100, with a range of average scores from 75 for the hardest marker to 85 for the easiest. Faculty members can be consistent with each other but exhibit systematic variation in average scores. This pattern is typical where faculty members can spot deviations from ideal but disagree over how serious these deviations are. The probability that students whose projects were evaluated by the faculty member who gave the lowest marks and just happened to be less talented was well less than $p<0.001$ in this study. Some students benefitted from the good fortune of having their typodonts picked up by easy graders.

### Component Criterion Consistency

Table 2 shows the summary of the results considering the detailed criterion subscores analysis of the ratings. In each matrix, the faculty rating is taken as the standard, and it is asked to what extent students agree with faculty members that a project is acceptable on each of the seven (denture set-up) or eighteen (immediate denture) criteria. On the first

**Note:** Confirmed Acceptable percentage is proportion of criteria on calibration cases marked acceptable by both the target faculty member and his or her colleagues. Confirmed Flawed percentage is proportion of criteria marked as flaws and confirmed by colleague. Self-reported easy grader is score on Likert item asking whether faculty member considers himself or herself to mark high (7) or low (1), with 4=average. Self-reported consistency is Likert item asking about consistency of marks given compared with faculty colleagues. Because of the small sample size, generally only $r$-values above 0.55 are statistically significant.
loading on that factor will perform well on other criteria loading on the same factor. It will not be possible to use that information, however, to predict performance on criteria that do not load on the same factor. A negative loading signifies a case in which students doing well on that criterion will tend to do poorly on the criterion with a positive loading. For example, Factor I of the denture set-up contains the “wax-up” criterion and the “ready for processing” criterion. Performance on these criteria is yoked.

**Table 2. Consistency of student and faculty member evaluation on specific criteria for denture set-up and immediate denture laboratory projects**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Acceptable</th>
<th>Flawed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denture Set-Up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty</td>
<td>2.098</td>
<td>1.580</td>
<td>3.679</td>
</tr>
<tr>
<td>Flawed</td>
<td>1.250</td>
<td>2.072</td>
<td>3.322</td>
</tr>
<tr>
<td>Total</td>
<td>3.348</td>
<td>3.652</td>
<td>7.000</td>
</tr>
<tr>
<td>Immediate Denture</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>12.898</td>
<td>2.454</td>
<td>15.362</td>
</tr>
<tr>
<td>Flawed</td>
<td>1.964</td>
<td>0.674</td>
<td>2.638</td>
</tr>
<tr>
<td>Total</td>
<td>14.862</td>
<td>3.138</td>
<td>18.000</td>
</tr>
<tr>
<td>Confirmed Acceptable</td>
<td>62.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confirmed Flawed</td>
<td>54.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>χ²=0.264</td>
<td>ϕ=0.194</td>
<td>r=0.262</td>
<td></td>
</tr>
<tr>
<td>Peer Immediate Denture</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faculty</td>
<td>12.607</td>
<td>2.230</td>
<td>14.807</td>
</tr>
<tr>
<td>Flawed</td>
<td>2.200</td>
<td>0.963</td>
<td>3.163</td>
</tr>
<tr>
<td>Total</td>
<td>14.807</td>
<td>3.193</td>
<td>18.000</td>
</tr>
<tr>
<td>Confirmed Acceptable</td>
<td>85.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confirmed Flawed</td>
<td>33.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>χ²=0.425</td>
<td>ϕ=0.154</td>
<td>r=0.184</td>
<td></td>
</tr>
</tbody>
</table>

**Latent Structure of Test Case Performance**

Tables 3 and 4 show results of the factor analyses. For clarity of presentation, only significant factor loadings are displayed in the tables. The denture set-up contains a clean, three-factor structure that accounts for 61 percent of the variance. Within each factor there is a statistically significant likelihood that those students who performed well on any criterion loading on that factor will perform well on other criteria loading on the same factor. It will not be possible to use that information, however, to predict performance on criteria that do not load on the same factor. A negative loading signifies a case in which students doing well on that criterion will tend to do poorly on the criterion with a positive loading. For example, Factor I of the denture set-up contains the “wax-up” criterion and the “ready for processing” criterion. Performance on these criteria is yoked.

**Table 3. Factor analysis of student performance on a denture set-up project in preclinical removable prosthodontics showing clusters of strong performance**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Points</th>
<th>Factor I</th>
<th>Factor II</th>
<th>Factor III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anterior arrangement</td>
<td>82%</td>
<td>0.616</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occlusal plane</td>
<td>87%</td>
<td>0.644</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ridge relationship</td>
<td>80%</td>
<td>0.727</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Centric</td>
<td>73%</td>
<td>0.745</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wax-up</td>
<td>70%</td>
<td>0.896</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Articulator settings</td>
<td>88%</td>
<td></td>
<td></td>
<td>-0.826</td>
</tr>
<tr>
<td>Readiness for processing</td>
<td>77%</td>
<td>0.860</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Points is the average “easiness” of each criterion: the number of points awarded by faculty members based on the number possible. For clarity, only statistically significant factor loadings are shown.
Factor III loads positively on “ridge relationship” and negatively on “articulator settings,” meaning that many students tended to perform well on one and poorly on the other. The structure for the immediate denture project had five significant factors. Although the five factors explained 48 percent of the variance, their interpretation is not as straightforward as with the denture set-up project.

The message from the factor analyses is that the criteria used for evaluation may be theoretically distinct but are not actually independent in practice. They cannot be summed to form an overall impression of the project without causing distortion. (Had there been no latent structure in these projects, there would have been a single, general factor with two ends of a continuum: “good student or poor student.”)

A “points” column is shown in Tables 3 and 4. This is the average number of possible points awarded by faculty members for each criterion. Higher scores indicate that most students performed well on that criterion, and low scores signified areas where instruction appears to have been deficient.

### Discussion

When viewed in the context of the literature, these findings prompt four points of discussion. First, it will be confirmed that checklist or component criterion evaluation, as an example of practice designed to promote “objectivity,” is not effective. Second, the purposes of evaluation as description or prediction will be contrasted. Then, efforts to save objectivity on other grounds will be examined. Finally, some advantages of using evaluation systems based on professional judgment will be presented.

### Problems with Component Criterion Evaluation

The most obvious finding in these data is that no “gold standard” appeared for removable prosthodontics laboratory projects such as denture set-ups or immediate dentures. More constructively, it may be possible to say that any objective ideal that might exist is subject to individual interpretation. If

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Points</th>
<th>Factor I</th>
<th>Factor II</th>
<th>Factor III</th>
<th>Factor IV</th>
<th>Factor V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mounting and articulator</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Casts center in articulator</td>
<td>93%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.800</td>
</tr>
<tr>
<td>Settings, incisal guide table</td>
<td>91%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vertical dimension</td>
<td>89%</td>
<td>0.399</td>
<td>-0.363</td>
<td></td>
<td>0.401</td>
<td></td>
</tr>
<tr>
<td>Anterior teeth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Position and appearance</td>
<td>86%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct length</td>
<td>89%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posterior teeth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stable centric</td>
<td>96%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.766</td>
</tr>
<tr>
<td>One stop per tooth</td>
<td>91%</td>
<td>-0.387</td>
<td>0.371</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buccal cusps, axial orientation</td>
<td>91%</td>
<td>0.451</td>
<td>-0.432</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grind-in</td>
<td>96%</td>
<td>0.831</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length and overlap</td>
<td>91%</td>
<td>0.736</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group function</td>
<td>73%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.671</td>
</tr>
<tr>
<td>No prematurities</td>
<td>80%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anatomy restored</td>
<td>96%</td>
<td>0.843</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Wax-up

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Points</th>
<th>Factor I</th>
<th>Factor II</th>
<th>Factor III</th>
<th>Factor IV</th>
<th>Factor V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wax bulk appropriate</td>
<td>74%</td>
<td></td>
<td></td>
<td></td>
<td>0.575</td>
<td></td>
</tr>
<tr>
<td>Smooth</td>
<td>83%</td>
<td>0.407</td>
<td>0.431</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform preformed palate</td>
<td>83%</td>
<td>0.506</td>
<td>0.372</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reproduced gingiva, rugae</td>
<td>65%</td>
<td>0.674</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clean, ready for processing</td>
<td>80%</td>
<td>0.380</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Points is the average “easiness” of each criterion; the number of points awarded by faculty members based on the number possible. For clarity, only statistically significant factor loadings are shown.
there were a single standard that did not depend on professional interpretation, faculty members would have agreed strongly with each other. This finding is consistent with previously published research.

Taking student work apart in hopes of finding objectivity in the pieces proved to be illusory as well. The factor analysis suggested the existence of latent structure in both projects, but these structures differed from the checklist criteria of the grading sheet. It is as if we wanted to measure physical fitness by weighing and measuring individuals. There is an association between the ratios of these measures, but they do not define fitness. Further, because there is a significant association among criteria that load on common factors, they cannot be considered independent. Summing across dimensions on a check-off sheet that are presumed to be independent, but are not necessarily so, introduces bias. It might be sound practice to divide questions on an essay test or stations on an OSCE among evaluators and then pool scores. But having faculty members individually assess criteria on a laboratory project that is in some sense an integral whole is likely to be a hodgepodge.

It might still be hoped that basing evaluations on component criteria, although less than perfect, will at least be preferable to overall professional judgment. Perhaps there is an advantage in guiding faculty members to focus on dimensions that are often spoken of when describing good work. The contrasting method would be to ask faculty members to use their professional judgment based on available evidence. To test whether objective component evaluation or comprehensive professional judgment is more consistent, we compared the phi-values (objective) with the r-values (professional judgment) as shown in Table 1. Of the four comparisons across two projects, all favored the use of professional judgment. For student-faculty comparison on the immediate denture (φ=0.089, r=0.407) and for faculty-faculty comparisons on the same project (φ=0.375, r=0.575), the differences were significant (p<0.01).

These comparisons refer only to the reliability (consistency) of component criterion and professional judgment evaluation. There is a natural test that can be applied to these data to gauge their relative predictive validity. The accuracy of faculty members using the component criteria process for the first project to predict performance on the second project was r=0.175. The course director used the professional judgment approach based on a single overall score rather than summing across components and was able to predict performance across the six-month period at the r=0.359 level. This is a statistically significant difference at p=0.02 using the test for difference in related correlations.

These findings are consistent with other reports in the literature, with the exception of the Goepfert and Kerber study that reported component grading to be more consistent than glance-and-grade. The difference is likely attributable to the design of the earlier report. Goepfert and Kerber had faculty members rate a set of typodonts for a pediatric restorative procedure twice, separated by a few weeks, using a general standard. Faculty members then underwent training using these same typodonts to develop a common set of criteria for the group and subsequently rated the same typodonts twice more using the criteria they developed for those typodonts. The consistency between the first two ratings was lower than the consistency of the ratings of the same typodonts following discussion on the third and fourth trials. Remarkably, the difference was not statistically significant.

**Objective Description versus Professional Prediction**

Of the two (objective components description or professional judgment), the evidence suggests that objectivity is the blunter instrument. But it might still be argued that this is what should be used because it is more useful in the contexts of dental practice and dental education. Pure description sounds more scientific, and if there are faults in the application of the descriptions, these, at least, are not to be blamed on the describers. This line of reasoning wants to say that even though objectivity may not be as reliable as judgment, at least it is more valid.

Reliability is about consistency of the scores themselves; validity is about the consistency of the scores with an external standard of interest. (The superior predictive capability of professional judgment over objective description mentioned above is an example of validity.) Description must be interpreted to be of value. It is assumed that students will understand what a score of 79 means and will act appropriately and that all will act in a consistent fashion and act as the instructor intends. It is assumed that graduation committees will understand what a D in clinical restorative dentistry in the final year means. Objective evaluation washes its hands of the interpretive context. It pushes off decisions about
dentistry and dental education to others. Objectivity
at the point of first contact does not ensure objectivity
along the chain of interpretation.

Professional judgment leans the other direc-
tion. It is future- and decision-oriented. Clinicians
diagnose for the sake of treatment. The features of
the patient they pay attention to are those that con-
nect existing conditions to improvements through
alternative courses of action. (To say that a dentist
diagnoses a Class II amalgam does not sound as
strange as, perhaps, it should.) The same should be
the case in dental education. What matters about
assessing performance is the potential it holds for
improving student performance. Commenting on
differences that do not make a difference may look
like consistency, but it is really just a waste of time.

One of the strong voices in the 1970s scene of
clinical evaluation was Gunnar Ryge and his widely
adopted R/S/T/V scoring system.44 The score cat-
egories were grounded in serviceability of the dental
product for future patient care. “Sierra” restorations
required no further comment or action. They were
restorations that would do what they were supposed
to do. “Tango” restorations were substandard as they
currently existed, but could be salvaged with a bit of
adjustment. “Victor” meant that the work would need
to be redone. “Romeo” was the category of superior
performance, with an operational definition some-
thing like needing to invite a friend over to have a
look. Ryge’s standard was anything but objective. It
pointed to what should be done with the restoration
in the future. It was action-oriented and thus pointed
toward validity.

An extension of the Ryge approach was used
at the University of the Pacific from about 1995 to
2010. A nine-point scale was employed to forecast
actions needed for students (rather than for teeth as
in Ryge’s model). Scores of 1 and 2 indicated that
the proper action would be to stop teaching the stu-
dent: there was every indication in the judgment of
the faculty member that the student lacked capacity
to become a dentist and that patients were at risk
being treated by such a student. Ratings of 3 and 4
indicated that the appropriate action would be to alter
the standard teaching program for the student because
the trajectory was pointing to eventual failure. These
scores flagged required remedial intervention. Scores
of 8 and 9 also indicated that the current educational
program should be terminated. They signaled compe-
tence: the student had absorbed all of the educational
benefit inherent in the current situation and needed
new challenges to ward off stagnation and Shenani-
gans and provide opportunities for further profes-
sional growth. The 5, 6, and 7 scores comprised the
vast majority of marks given, and faculty members
often agonized over the distinctions. Nothing differ-
ent happened to any of these students as a result of
receiving marks anywhere in this range, and 5s were
treated the same as 7s. These marks signaled only
that the student was progressing satisfactorily in the
common educational program and should continue.

Research reports measuring consistency of
evaluations of the predicted clinical serviceability of
restorations tend to be higher than consistency of ob-
jective feature descriptions of those restorations.45-50
Field reports of the Ryge system for evaluating the
future usability of restorations and the competency
evaluation system for evaluating the future career
paths of students claimed internal consistency in the
0.700 < r < 0.850 range and predictive validity
coefficients above r=0.500.51-53

The “points” columns in Tables 3 and 4 con-
tain useful information about the relative strengths
and weaknesses of the instructors and the program
for bringing students to standard. On the immediate
denture project, there were no students who had flaws
on more than one-quarter of the criteria. However,
more than 25 percent of students were deficient on
“group function” and “bulk of wax”—despite turn-
ing in overall satisfactory projects. One-third of the
students were below par on “reproduction of gingiva
and rugae.” These are not student problems; they are
problems with the educational program.

Objectivity Anyway

Despite having somewhat tarnished credibility
as a useful foundation in dental education, we may
still want to ground student evaluation in “objective”
description for several reasons. There is a belief about
the relationship between reliability and validity that
has historically operated to pull attention away from
validity because reliability is thought to be in a prior
position. The legal liability associated with student
evaluation is never far from the minds of academic
administrators. The best argument, however, for
objectivity is probably that it furnishes easy cover
for faculty members who are bothers by having to
justify their decisions.

The psychometric argument. It is sometimes
said that “objectivity” is a baseline requirement for
evaluation. This is expressed in the saying “there can
be no validity without reliability.” This is true, but
professional judgment is more appropriate for improving opinions about professional education for the judgment of those who are qualified by training to make these decisions—i.e., faculty members.

This case also established the standards for “arbitrary” and “capricious” judgments and the powerful role of due process. An arbitrary action is one that does not have a sound justification. Basing academic decisions on objective scores with low interfaculty consistency and no documentation of predictive validity would seem to be arbitrary in this sense. A capricious action is one that is applied differently to different students or in different cases. This is just basic fairness. The due process clause is perhaps the most important in the set. This is where administrators and faculty members in dental school may butt heads. Faculty members could be right in saying that a given student should not be a dentist, but they have to give the administration something to stand on in case the matter is challenged in court.

Professional judgment normally meets the legal standards of not being arbitrary or capricious and fits due process when normal disciplinary procedures are followed. Objectivity may work as well, although there are some dangers in the direction of being arbitrary.

The indisputable description argument. Claims of expediency for objective grading do not appear in the literature. If objectivity is really indisputable description or warranted assertability, that marks a natural place to end discussion. As such, objective pronouncements protect faculty members from having to defend their judgments.

The use of checklists certainly makes it possible for novice faculty members to participate in evaluating complex student performance. This approach saves time. Evaluations reduced to numbers facilitate averaging and other manipulations that can be done by staff. They also distance performance from decision makers, with the impression that such a gap does not require being filled.

All of these features of objectivity seem to be accurate descriptions.

Professional Judgment Instead

Faced with a choice between believing that “objectivity” is really there but we are poor at getting our hands on it and the alternative that professional judgment is more appropriate for improving
practice and teaching, dental education has tended to, at least in public, honor the prospect of objectivity and privately get on with judgments. Building a better case for objectivity in the face of limited consistency among evaluators has taken various forms. Calibration exercises and use of component criteria are occasionally attempted, but there is no evidence that these produce large or lasting benefits. For the most part, the problem is simply ignored. Three other approaches have strong records in the psychometric literature and can be added onto existing preclinical and clinical practices without having to change human nature.

**Guaranteeing a target consistency.** Given some actual general level of consistency among faculty evaluators, it is possible to achieve a satisfactory standard of consistency without having to change the faculty members. If the course director in the preclinical removable prosthodontics course described here had wanted consistency in scoring to be \( r = 0.80 \) rather than \( r = 0.57 \), that could have been achieved by having each project marked by three faculty members instead of one. No change would be needed in selection of faculty, calibration or training, or component criteria standards. Seven faculty members per project would be needed to get to \( r = 0.90 \). The NBDE has such high reliability due in large part to the fact that the tests are so long. The easiest way to improve the reliability of a multiple-choice test (and thus the easiest way to avoid getting it wrong about which students know enough to progress in the program) is to increase the length of the test.

The formula for determining whether it is worth the additional cost to be more accurate is called the Spearman-Brown Prophecy Formula. It is easy to calculate: Increase in Information = \[ \text{Desired Reliability} \times (1 – \text{Actual Reliability}) \] / \[ \text{Actual Reliability} \times (1 – \text{Desired Reliability}) \]. Increase in Information is a small number, such as 2.5 or 6, that identifies what the current amount of information needs to be increased by to get the desired effect. 2.5 would mean that the number of examiners or test questions should be multiplied by 2.5; 6 would mean that six times as many units containing the same information need to be involved. Accuracy in information on which decisions are made about procedures or students comes with a cost. The Spearman-Brown formula is a simple way of estimating that cost.

**Protecting against bias.** Consistency across faculty evaluators reflects reliability, but not bias. Imagine two faculty members: one rates three projects 4, 5, and 6, while the other judges the same projects 5, 6, and 7. The correlation for reliability is a perfect \( r = 1.00 \), but every student would want to be evaluated by the second faculty member.

Rather than train, calibrate, or deselect faculty members who maintain private standards that seem objective to them, it is easier and less costly to standardize the scores. The assumption is that each faculty member will encounter, by chance and in the long run, students whose scores will be similar on average. If that is so, the average score for each faculty member should be the same if we can find a good correction for personal bias. The formula works like this: \([\frac{(\text{Score} – \text{Faculty Average})}{\text{Faculty Standard Deviation}}] + \text{Group Average}\). Each student is judged against the standards of the instructor he or she happens to inherit, and the score is re-expressed in terms of the way the individual faculty member’s scores relate to the average of his or her peers. The scores are standardized even when the evaluators are not. A student who is average for Evaluator A is set to be average for all evaluators. This correction can be programmed into a spreadsheet and calculated with no additional effort by students, faculty members, or course directors.

**Study club model.** The third suggestion flows from the fact that almost all evaluation information is wasted. Objective descriptions depend on others taking appropriate action, even if they do not happen to understand or agree with the “objectivity” of the evaluator. In most cases, differences arrived at objectively will not have any effect on future behavior. What we need is the opportunity to convert evaluation programs into activities that influence future behavior, especially learning by students and faculty members.

A program that might serve this purpose would be for teams of students and faculty members to critique projects together. The questions that form the basis for this discussion are whether the students are aware of performance that could be improved and whether they know how to make those improvements. This discussion can result in plans for improved performance or, in rare cases, discussions about additional or different approaches to learning or even recognition of severe barriers such as dyslexia, domestic distractions, disinterest in the profession, or lack of capacity so great that dentistry should be abandoned. The only marks useful from this activity that need to be recorded for use by others are the four categories described in the competency grading program outlined above: consider ending dental education, participate in
remedial work, keep going as planned, or be given advanced or enriched opportunities. These category determinations are not averaged (two “remediations” and three “enrichments” do not equal to “do nothing”). A threshold number of marks in any category other than “continue as planned” would be sufficient to trigger action.

These discussion sessions should take place shortly after the project is finished and should be conducted in groups of two faculty members and two students. The reason for multiple participants is so that a range of information can be shared and to develop a better understanding of what students and faculty need to do in the future. Pairs of faculty members are needed to avoid the appearance that any one faculty member has the gold standard and to show that faculty members can continue to learn about both evaluation calibration and needed instructional emphasis. Student understanding will grow as well.

It may be objected that students should not see each other’s work in a setting where comparisons could be made. It might even be objected that faculty members should be free from the responsibility of having to defend the marks they give in the presence of their colleagues. These are small views of professionals. Increasingly, dental schools are using cross-class teams to deliver clinical care. In these settings, students not only see each other’s work; they cooperate on treatment plans that include, among other things, the evaluation of the serviceability of work that has been performed to that point. Practitioners regularly see each other’s dentistry by referral or when patients change dentists. Based on recent “Ethical Moment” columns in the Journal of the American Dental Association, dentists are not especially skilled in commenting on the performance of their colleagues. Such group comparison of actual projects is the essence of the study club model where the goal is to build professional judgment rather than offer objective descriptions.

Conclusion

In this study, an analysis of two removable prosthodontics technique projects in one U.S. dental school found that the use of component criteria (checklist) grading was less consistent than overall judgments of the same work and less predictive of dental students’ future learning. A factor analysis revealed latent structures in both projects that would make it inappropriate to use a component criteria approach for grading. In this article, we questioned common defenses of objectivity such as scientific foundation, the relationship between reliability and validity, and legal requirements. We also demonstrated how simple adjustments to judgment scores can be made more effectively than using checklists, faculty calibration, or deselecting faculty members and with better measurement and teaching features.

REFERENCES


